A black background with white text

AI-generated content may be incorrect.

Bachelor Thesis

**Income Estimation Models Analysis: Balancing Accuracy with Regulatory Requirements in the Georgian Banking Sector**

**Sandro Gogaladze**

*Bachelor’s Project Report is submitted for the Economics Program of International School of Economics at Tbilisi State University*

Project Supervisor: Shota Natenadze

**Tbilisi, 2025**

# Table of Contents

[Table of Contents ii](#_Toc198488645)

[List of Abbreviation iii](#_Toc198488646)

[Introduction 1](#_Toc198488647)

[Literature Review 2](#_Toc198488648)

[Empirical Foundation 3](#_Toc198488649)

[Data 3](#_Toc198488650)

[Preprocessing Strategy 4](#_Toc198488651)

[Methodology 6](#_Toc198488652)

[Commercial vs. Regulatory Objectives 6](#_Toc198488653)

[Approaches to Regulation-Aware Modeling 7](#_Toc198488654)

[Models 8](#_Toc198488655)

[Baseline Xgboost Model 8](#_Toc198488656)

[Model Selection and Design 8](#_Toc198488657)

[Performance Evaluation 9](#_Toc198488658)

[Post-hoc Calibration 12](#_Toc198488659)

[Quantile Regression as a Calibration Tool 12](#_Toc198488660)

[Pre-hoc Adjusting 16](#_Toc198488661)

[Huber Loss with Dynamic Threshold Penalty 16](#_Toc198488662)

[Segment-Aware Huber with Threshold Penalty 19](#_Toc198488663)

[Comparison of Pre-Hoc Objective Functions 23](#_Toc198488664)

[Limitations and Directions for Future Work 24](#_Toc198488665)

[Bibliography 26](#_Toc198488666)

[Appendix A: Computational Source – Jupyter Notebook 27](#_Toc198488667)

# List of Abbreviation

BOG Bank of Georgia

CTGAN Conditional Tabular Generative Adversarial Network

GEL Georgian Lari (currency)

KNN K-Nearest Neighbors

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

NBG National Bank of Georgia

PDP Partial Dependence Plot

PTI Payment-to-Income (Ratio)

R² Coefficient of Determination

RF Random Forest

RMSE Root Mean Squared Error

SDV Synthetic Data Vault

SHAP SHapley Additive exPlanations

XGBoost Extreme Gradient Boosting

# Introduction

Accurately assessing a borrower's creditworthiness is an essential part of the lending process for banks and other financial institutions. Prior to extending credit, banks have to verify an applicant's income to comply with regulatory standards and to evaluate the applicant’s ability to meet future repayment obligations. Verified income is also used to derive risk metrics, such as the Payment-to-Income (PTI) ratio, which measures the share of income allocated to debt servicing. This information is important for determining appropriate loan terms and ensuring that lending decisions are both responsible and risk-informed.

Traditionally, income verification involves manual and highly document-intensive processes, requiring financial institutions to carefully review pay stubs, tax documents, revenue service statements, bank statements, and employer references. This approach, although thorough, is inherently lengthy, cumbersome, and inefficient, often resulting in prolonged loan approval times and increased operational expenses for banks. In addition, individuals without officially verified incomes — such as self-employed entrepreneurs, freelancers, street vendors, independent contractors, gig economy workers, small-scale farmers, etc. — face significant disadvantages, as they frequently receive payments directly in cash, making it challenging or impossible to demonstrate their true earnings formally. Even borrowers whose incomes are recorded by revenue services commonly under-disclose their actual incomes, typically due to supplementary cash payments. Therefore, traditional verification methods often yield incomplete or inaccurate assessments, proving ineffective both for financial institutions, which incur higher costs and operational burdens, and for borrowers, who experience barriers or delays when attempting to access timely credit.

Due to increased competition, financial innovation, and a growing emphasis on operational efficiency, banks are moving away from traditional income verification methods towards automated income estimation models that employ various data sources and advanced analytic methods to approximate borrower incomes more efficiently.

This thesis analyzes the adoption and effectiveness of such income estimation models within the specific context of Georgia. Recently, the National Bank of Georgia (NBG) granted financial institutions the regulatory authority to implement statistical, machine learning, and artificial intelligence-based methods for income verification (National Bank of Georgia, 2020). Many Georgian banks have embraced these opportunities, with some already applying these models in practice and others actively developing and refining their analytical capabilities. Given the active interest and inherent complexity of this topic, this thesis explores and evaluates various modeling approaches, comparing their methodologies and performance concerning both predictive accuracy and regulatory compliance within the Georgian financial environment.

# Literature Review

Income prediction is a relatively new area of research, and historical academic literature on the topic is limited. However, there has been increased interest in the past several years with researchers working on modeling various strategies.

Kibekbaev and Duman (2016) recently conducted large-scale benchmarking studies on income prediction using real-life datasets from Turkish banks. Their research compared 16 regression algorithms, including OLS, Ridge, Robust Regression, MARS, ANN, LS-SVM, and several two-stage models. They found that while non-linear and two-stage models outperformed others, traditional linear regression models still delivered competitive results, suggesting that even simple models can yield valuable insights when applied thoughtfully.

Building on this work, Sargan (2021) introduced an ensemble-based income prediction model combining XGBoost, LightGBM, and Random Forest. Using Turkish banking data, the study showed that segmenting customers by education level and including behavioral and regulatory features significantly improved model accuracy and credit risk assessment.

Matkowski (2021) developed an interpretable income prediction framework for bank clients using the eXtreme Gradient Boosting (XGBoost) algorithm in conjunction with SHAP (SHapley Additive exPlanations). Using a real-world dataset from a bank in Poland, the author focused on the model's predictive ability and the interpretability aspect of the model, which is an important factor in many regulated domains.

One of the earliest works in the field is Lazar (2004), who applied Support Vector Machines (SVM) to income prediction using the U.S. Census Bureau’s Current Population Survey data. Even though this study was framed as a classification problem - classifying whether an individual's salary was more than $50K - it provides useful insights about computational tradeoffs and feature selection. This study illustrates how early ML approaches were able to extract useful signals about income from real demographic and occupational data, despite computational limitations at the time.

The JPMorgan Chase Institute (2018) introduced a machine learning-based model for estimating gross family income using administrative banking data. The study found that relying solely on checking account inflows does not provide an accurate picture of a customer’s income. However, when these inflows are analyzed alongside other indicators—such as credit card limits, total liquid assets, and neighborhood-level data like ZIP code characteristics—the accuracy of income estimation improves considerably.

Similar to the JPMorgan study, Suárez et al. (2021) demonstrated that alternate data sources can significantly improve income prediction models. The authors developed an XGBoost model utilising behavioural and transactional data from a Super-App, demonstrating substantial predictive capability even in the absence of bureau data. The study emphasised that alternative data not only improves model accuracy, but also facilitates income calculation for previously unbanked or financially invisible populations—a critical advancement towards greater financial inclusion.

Overall, these studies suggest a clear global trend, towards the automation of estimating income using machine learning and alternative data. Therefore, it is valuable to examine how these approaches can be adapted to the local context. This thesis fills the gap by focusing on income prediction models within the Georgian banking sector.

# Empirical Foundation

## Data

In order to explore and analyze income estimation models, access to relevant, high-quality data is essential. Income prediction typically relies on transactional, credit-related, and sometimes demographic information. These data are exclusively held by, and available to, financial institutions. Since this type of data is personal and sensitive, legal and regulatory restrictions prevent public sharing. Therefore, this thesis is carried out in collaboration with the Bank of Georgia (BOG), the leading financial organisation in Georgia. For research purposes, BOG provided a synthetic dataset designed to replicate the structure, distribution, and statistical properties of real customer data while ensuring individual privacy is preserved.

The synthetic dataset was created using the CTGANSynthesizer method from the Synthetic Data Vault (SDV) library. CTGAN (Conditional Tabular Generative Adversarial Network) is an advanced deep generative model specifically created to address the problems of tabular data such as class imbalance, mixed data types, and complex nonlinear relationships. In contrast with classical over-sampling or noise-based approaches, CTGAN learns the joint distribution of the features and can generate highly realistic synthetic samples that retain the dependencies observed in real data. This makes it well-suited for this financial dataset where both privacy and structural fidelity is critical.

The resulting dataset contains 21 features and a target variable representing the client’s stable monthly income. Unlike temporary inflows such as bonuses or recent salary deposits, the target reflects a consistent, recurring income level. This ensures the model is trained to estimate the client’s long-term financial capacity rather than short-term or irregular transactions.

Feature selection was guided by expert judgement from BOG, prioritizing variables that reflect transactional behavior, account balances, and credit registry data. The feature set includes past and current income indicators (Inc\_Past, Inc\_6M, Inc\_Past\_avg, Inc\_Past\_max), liabilities (Liab\_Tot), transfer activity (Transfers\_in, Transfers\_out, Min\_transfer\_In, Acct\_Trns), account usage (Turnover, Transactions, Tot\_in, Balance, Bal\_Cur), and loan-related metrics (Loan, Loan\_Cnt, Payments, Payment\_L). Demographic variables were intentionally excluded, as they offered limited predictive value. This focused selection ensures the model remains interpretable, efficient, and practical for deployment.

## Preprocessing Strategy

During the development phase, multiple preprocessing strategies were tested to improve data quality and model performance. After empirical evaluation, a streamlined and robust pipeline was adopted as the final preprocessing strategy, balancing effectiveness, interpretability, and deployability.

The first step in the pipeline addresses missing values using the K-Nearest Neighbors (KNN) imputation method. As no systematic patterns of missingness was identified during exploratory analysis, the missing observations was assumed to be randomly distributed across the dataset. KNN imputation estimates missing values by averaging the values of the most similar observations, based on feature proximity. This approach preserves the local structure of the data and avoids the distortions that can arise from global methods such as mean or median imputation, which often overlook contextual relationships between variables.

To mitigate the influence of extreme values, outlier capping was applied using feature-specific strategies based on the distributional characteristics of each variable. Features exhibiting significant skewness or long upper tails were capped using the interquartile range or robust Z-score methods, while more stable but business-critical features were treated using percentile-based thresholds. This approach ensures that outliers do not disproportionately influence model training, while retaining sufficient variance for predictive signal.

In the final stage of preprocessing, new features were engineered by constructing interpretable financial ratios that normalise key quantities and highlight behavioral patterns. Examples include the loan-to-income ratio, which measures a borrower's debt burden relative to income; the balance-to-liability ratio, which reflects liquidity; and income growth indicators based on historical inflow comparisons. These derived features enhance the model’s ability to generalize across diverse borrower profiles and income levels.

The preprocessing pipeline was applied separately to the training and test subsets, ensuring no data leakage and maintaining the integrity of model evaluation. The resulting datasets are clean, structured, and suitable for predictive modeling in a regulatory and operational banking environment.

# Methodology

## Commercial vs. Regulatory Objectives

A primary challenge in income estimation modeling is defining objectives that balance institutional goals with regulatory expectations. Commercial banks typically seek to minimize overall prediction error to enhance credit assessments and operational efficiency. In contrast, regulatory authorities—such as the National Bank of Georgia (NBG)—focus more on limiting income overestimation to ensure financial system stability. These differing priorities create a fundamental trade-off between predictive accuracy and regulatory conservatism.

According to insights gathered through interviews and discussions with NBG representatives, current supervisory practices in Georgia tolerate some level of overestimation. While thresholds may vary by context, it is generally accepted that predictions may exceed actual income by up to 10%, provided that such cases are limited to no more than 10% of all observations. These parameters are not fixed but serve as practical benchmarks for model evaluation.

In this thesis, a more relaxed but conceptually consistent framework is adopted to allow for greater flexibility and experimentation during model development. Specifically, a prediction is considered acceptable if it does not exceed the true income by more than 200 GEL or 20 percent—whichever is greater:

To quantify compliance, a binary violation indicator is introduced:

The model is considered compliant if the proportion of violating predictions remains below 10%:

This dual-threshold — or dynamic threshold — approach plays a crucial role in addressing the limitations of percentage-based error metrics. In low-income segments, even small absolute differences can appear disproportionately large when expressed as percentages, which may give a misleading impression of model performance. Adding a fixed GEL-based threshold helps solve this issue by setting a minimum margin that accounts for normal variation. The 200 GEL floor was chosen based on domain knowledge and input from supervisors, as it provides a reasonable allowance—especially for clients earning under 1,500 GEL—where such deviations are relatively common and considered acceptable in real-world evaluations. This combination ensures the evaluation framework remains fair, practical, and aligned with supervisory expectations across different income levels.

To assess the accuracy of the models from both the institutional and regulatory perspectives, this thesis reports on two sets of metrics. Financial institutions largely focus on traditional statistical measures - such as Mean Absolute Error (MAE), Coefficient of Determination (R²) - as acceptable indications of predictive accuracy or operational efficiency. In comparison, regulators apply compliance-based indicators especially frequency and severity of thresholds breached. These overestimation-related metrics provide a more conservative and risk-aware view of model behavior.

To ensure fairness, consistent treatment, and to mitigate hidden biases, performance is also reported separately across borrower income segments. In this thesis, borrowers are categorized into three income groups based on equal quantiles to ensure balanced representation: low-income (<1500 GEL), middle-income (1500–2500 GEL), and high-income (>2500 GEL).

## Approaches to Regulation-Aware Modeling

To address both institutional performance objectives and regulatory constraints, this thesis investigates two categories of corrective strategies: post-hoc and pre-hoc.

Post-hoc methods are applied after a model has been trained and are especially useful when the model performs well overall but produces outputs that exceed regulatory thresholds. These approaches are commonly used when retraining the model is impractical, resource-intensive, or undesirable. Techniques such as **prediction clipping, quantile-based calibration**, and **output scaling** are designed to reduce the risk of income overestimation without modifying the model’s internal structure. These post-processing layers are also widely accepted by regulators. As confirmed through interviews conducted for this thesis, the **National Bank of Georgia (NBG)** may permit the use of statistical or machine learning models **on the condition that all predictions are passed through a compliant post-hoc adjustment mechanism**. In some cases, supervisors may derive and determine this adjustment layer themselves. This regulatory safeguard ensures that final outputs stay within acceptable supervisory boundaries and helps mitigate systemic risk — even when the original model occasionally produces overestimated predictions.

Pre-hoc methods, on the other hand, embed regulatory requirements directly into the model training process. This is typically achieved by altering the objective function to penalize threshold violations or by introducing constraints during optimization. Such methods encourage the model to internalize the cost of overestimation, making them more suitable for applications in strictly regulated domains where conservatism is not optional but foundational.

# Models

## Baseline Xgboost Model

### **Model Selection and Design**

The first modeling attempt focused solely on maximizing predictive accuracy. Given the complexity of income estimation — marked by non-linear relationships, diverse transaction behaviors, and significant individual heterogeneity — tree-based machine learning models were identified as suitable candidates for baseline evaluation. Their ability to model complex patterns and handle heterogeneous financial data with minimal preprocessing makes them potentially well-suited to this task. Based on these considerations, a tree-based model was selected as the starting point for further experimentation.

As part of the model selection process, several tree-based regression algorithms were evaluated, including XGBoost, LightGBM, CatBoost, and Random Forest. All models exhibited broadly comparable performance during testing. XGBoost was ultimately selected as the baseline model — not due to clear predictive superiority, but primarily for practical reasons such as computational efficiency, widespread adoption in structured data applications, and strong compatibility with established interpretability tools.

To optimize model performance, a randomized hyperparameter search with five-fold cross-validation was conducted. The parameter space included values for the number of trees, tree depth, learning rate, and regularization terms. The objective function minimized root mean squared error (RMSE), balancing predictive power with generalization capability. The resulting model was trained on the entire training set using the best-performing configuration.

### **Performance Evaluation**

The baseline model’s performance was evaluated on a held-out test set using both standard regression metrics and regulation-specific indicators. Its results closely mirrored the training outcomes, indicating a well-generalized model with no signs of underfitting. This consistency suggests that the model learned stable relationships and performs reliably under traditional statistical benchmarks.

From a conventional modeling perspective, the results were encouraging. The model achieved a mean absolute error (MAE) of approximately 651 and a coefficient of determination (R²) of 0.75, indicating that it captured a substantial portion of income variability across the population. However, when assessed through a regulatory lens, significant shortcomings became evident.

Although aggregate performance was acceptable, the model frequently failed to meet supervisory standards. Only 24.5% of test set predictions fell within ±10% of the true income, and just 46.1% were within ±20%. More critically, 33.6% of the predictions exceeded the regulatory compliance threshold—defined as more than 200 GEL or 20% above the actual income. This violation rate significantly exceeds the acceptable limit of 10%, highlighting the model's misalignment with regulatory requirements.

These regulatory violations were especially pronounced among lower-income individuals. In the low-income segment (<1500 GEL), around 57% of predictions breached the compliance threshold, highlighting a high risk of overestimation. In contrast, performance improved in the high-income group (>2500 GEL), where the threshold breach rate fell to under 14%. These findings highlight a key insight: the behavioral and financial differences across income segments are substantial, making it inherently difficult for a single model to perform well across all groups. The sharp disparity in prediction accuracy and compliance—especially the severe overestimation in the low-income segment—demonstrates the limitations of a uniform approach. This reinforces the need for segment-aware modeling strategies, which are explored later in the thesis.

Figure 1 visualizes these outcomes. The calibration plot (top-left) shows that predictions are well-aligned with actual values in higher-income groups but increasingly diverge in lower ranges. The error histogram (top-right) reflects balanced overall bias, but this masks large relative errors concentrated in low-income cases. Threshold violation plots (bottom-right) confirm that the regulatory breaches are not randomly distributed but are most severe in segments where conservative prediction is most crucial.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 1: Baseline Model Prediction Analysis Across the Test Set.

To better understand the drivers behind these predictions, SHAP (SHapley Additive exPlanations) values were used to interpret the model. SHAP decomposes each prediction into feature-level contributions, making it possible to explain not only which features were most influential overall, but also how they affected specific predictions—an essential property in supervised financial modeling.

As shown in Figure 2, the most impactful features were Min\_transfer\_In, Inc\_in, and Turnover, all of which reflect short-term inflow activity. These results are intuitively reasonable: income estimation should naturally be influenced by recent deposit behavior and transaction volume. The model’s emphasis on such features aligns well with domain knowledge, where active account usage and consistent inflows are generally strong indicators of financial capacity.

A graph with blue and pink lines

AI-generated content may be incorrect.

Figure 2: SHAP summary plot showing feature contributions to the baseline model’s predictions.

However, despite appearing reasonable, the model’s reliance on short-term inflows introduces a critical vulnerability. It did not differentiate between the source or stability of transactions—treating all deposits, whether recurring salaries or one-time transfers, as equally informative. This led to frequent overestimations, particularly where temporary liquidity was mistaken for sustainable income in the absence of consistent historical inflow patterns.

Local interpretation confirms this behavior. High-error cases often exhibited extreme Loan-to-Income or Balance-to-Liability ratios, indicating financial inconsistencies that the model failed to penalize. Although the dataset contained a wide range of transactional and credit indicators, these captured only aggregate activity without contextual detail. As a result, the model frequently misinterpreted temporary spikes in cash flow as evidence of long-term financial stability.

Another contributing factor is the nature of the data. The analysis was conducted using synthetically generated data, designed to replicate the statistical characteristics of real customers. While effective for preserving privacy and enabling experimentation, it may not fully capture rare edge cases or behavioral outliers. During development, the synthetic dataset was treated as a functional proxy for real-world data—an assumption that was necessary for experimentation but may have introduced distortions in the model’s results.

In summary, while the baseline XGBoost model performs well under standard metrics and generalizes reliably in aggregate, it fails to meet key regulatory requirements—particularly in low-income segment, which is more vulnerable to the consequences of overestimation. These findings show that the default configuration is not suitable for supervised lending decisions. The following sections introduce post-hoc and pre-hoc strategies designed to align the model with institutional goals and regulatory expectations.

## Post-hoc Calibration

### Quantile Regression as a Calibration Tool

To address the shortcomings of the baseline model—particularly its failure to meet regulatory constraints—a post-hoc calibration technique can be applied to adjust the model’s outputs without retraining or altering its internal structure. In this thesis, a quantile regression-based calibration method was explored as a way to conservatively shift predictions downward and reduce the risk of income overestimation.

Quantile regression estimates a specific percentile of the true income distribution conditioned on the model’s original prediction. This allows the adjusted output to be intentionally conservative—ensuring that a fraction  of actual incomes are expected to meet or exceed the calibrated prediction.

The adjustment is implemented through a simple linear transformation of the original predictions, expressed as:

The coefficients and​ are estimated by minimizing the quantile loss function:

where is the check function defined as:

This asymmetric loss penalizes overestimations more heavily than underestimations when  For example, at , overestimations are penalized four times more than underestimations of equal magnitude. As a result, the model learns to shift predictions downward, reducing the likelihood of excessive income estimates.

While this approach does not directly enforce the regulatory threshold—defined above as predictions exceeding either 200 GEL or 20 percent above the actual income—it serves as a close empirical approximation. The quantile level  was selected through experimentation on the training set, where several candidate values were tested to identify the one that brought the violation rate close to the 10% benchmark. The best-performing quantile, , was then applied to the test set for final evaluation. To ensure the validity of the results and avoid data leakage, all calibration procedures maintained a strict separation between training and test data.

**Results**

Using the selected quantile level  the quantile regression calibration produced the following linear adjustment function:

This transformation introduced a deliberate downward shift in predictions. As a result, the calibrated model reduced threshold violations from 33.6% to 9.9%, successfully aligning with the regulatory limit at the aggregate level. Figure 3 provides a detailed view of this outcome, illustrating the tightened error distribution and the overall shift toward more conservative predictions. The calibration reduced severe overestimations and increased underestimations, ultimately improving compliance with supervisory expectations.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 3: Test set prediction analysis after quantile calibration (q = 0.20).

However, the improvement was not uniform across all income groups. As shown in Figure 4, while violation rates dropped significantly for middle- and high-income borrowers, the low-income segment (≤ 1500 GEL) continued to exceed the 10% threshold (bottom-right plot). This highlights a key limitation of the fixed quantile approach: it applies the same adjustment across the board, without accounting for heterogeneity in income levels or error sensitivity. As a result, the method achieves compliance at the aggregate level but remains inconsistent across borrower segments.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 4: Effects of quantile calibration (q = 0.20) on predictions, error distributions, and threshold exceedance by income group.

These regulatory gains also came at the cost of predictive accuracy. The Mean Absolute Error (MAE) increased from 651.03 to 801.00, and the Root Mean Squared Error (RMSE) rose due to the broad downward bias introduced by the calibration. The R-squared score dropped from 0.75 to 0.62, indicating a reduced ability to explain variation in actual incomes. This outcome illustrates a central trade-off in conservative post-hoc calibration: regulatory robustness is improved, but at the expense of overall model precision.

To overcome these limitations, the next section explores a pre-hoc calibration strategy that integrates regulatory priorities directly into model training—enabling more adaptive, segment-aware prediction behavior from the outset.

## Pre-hoc Adjusting

To address the shortcomings of post-hoc calibration, a more effective approach involves modifying the objective function during training. This pre-hoc strategy enables the model to incorporate regulatory constraints directly into the learning process, rather than relying on adjustments after predictions are made.

In this framework, the loss function is extended to account for both predictive accuracy and supervisory compliance. The model is penalized not only for large errors, but also for overestimations that exceed borrower-specific regulatory thresholds. This encourages conservative behavior during training, aligning the model’s outputs with the criteria on which it will be evaluated.

Several custom objective functions were explored during experimentation. Among them, two formulations proved most effective. Both build on a robust Huber loss foundation and incorporate penalty terms for threshold exceedance to discourage non-compliant predictions during model training.

### Huber Loss with Dynamic Threshold Penalty

The first objective function extends the standard Huber loss by adding a squared penalty for predictions that exceed a borrower-specific regulatory threshold. This formulation encourages the model to remain accurate while avoiding overestimations that would violate compliance constraints.

Huber loss was chosen as the base component due to its robustness in handling noisy financial data. Unlike Mean Squared Error (MSE), which heavily penalizes large deviations and is prone to overfitting outliers, Huber loss transitions to a linear penalty beyond a threshold , reducing the influence of extreme values. This is particularly important for income prediction, as demonstrated by the Georgian dataset, which features irregular account activity with high inflow spikes—often driven by informal or non-recurring transactions. These distortions are especially common among low-income borrowers and can mislead the model if not properly controlled. The Huber framework helps mitigate this issue, while the added penalty term enforces compliance by discouraging excessive overestimations.

**Loss Function Formulation**

Formally, the Huber loss is defined as:

To align the model with regulatory expectations, this base loss was augmented with an additional penalty that activates only when the predicted income  exceeds the borrower-specific threshold .

The full loss function becomes:

where:

* defines the regulatory threshold,
* is the transition point in the Huber loss that separates small and large errors,
* controls the strength of the penalty for threshold exceedance.

The penalty weight was selected based on experimentation with training performance. Lower values failed to reduce regulatory violations, while higher values led to excessive underestimation. All tuning was performed on the training set to avoid data leakage, and final evaluation was conducted using a separate test set.

**Results**

The model trained with the custom “huber\_plus\_threshold\_loss“ objective was evaluated over 300 boosting rounds. Training progressed without overfitting, and the final performance on the test set yielded a Root Mean Squared Error (RMSE) of 1241.62 and a Mean Absolute Error (MAE) of 782.56. While these figures indicate a reduction in accuracy compared to the baseline, the trade-off was expected given the introduction of constraints focused on reducing overestimation.

The most significant improvement was observed in regulatory alignment. The threshold exceedance rate dropped from 33.56% to 10.47%, bringing the model’s predictions in line with the compliance target. This demonstrates that integrating penalties directly into the learning process can effectively enforce conservative prediction behavior without the need for post-hoc correction mechanisms.

Figure 5 illustrates how the pre-hoc model compares to the baseline and post-hoc alternatives. Unlike the post-hoc quantile approach—which achieves compliance by applying a sharp downward shift to most predictions—the pre-hoc model achieves similar regulatory outcomes with less distortion. It adjusts predictions more selectively, resulting in fewer cases of excessive underestimation and preserving closer alignment with actual incomes.

At the same time, Figure 5 (bottom-left) reveals a limitation. Although the overall threshold violation rate improved, compliance gains were not evenly distributed across income segments. The low-income group, in particular, still exhibited a violation rate of around 22%, compared to less than 10% in the mid- and high-income segments. This result suggests that the model’s penalty mechanism, while effective at the aggregate level, lacks the flexibility to adjust to differences in borrower risk and income volatility.

A collage of graphs and diagrams

AI-generated content may be incorrect.

Figure 5: Comparison of baseline, post-hoc, and pre-hoc models across prediction error, threshold exceedance, and calibration.

To ensure that the model remained interpretable despite the additional constraints, SHAP values were used to examine feature importance (Figure 6). The most influential variables—such as Min\_transfer\_In, Inc\_6M, and Payment\_L—remained consistent with domain expectations, emphasizing transaction inflows and repayment behavior. The SHAP summary plot indicates stable and monotonic relationships, confirming that the modified objective did not compromise transparency or introduce erratic behavior.

A graph with red and blue lines

AI-generated content may be incorrect.

Figure 6: SHAP summary plot showing feature contributions to the pre-hoc model’s predictions.

Overall, the pre-hoc model presents a well-balanced improvement over previous strategies. It meets regulatory expectations while preserving interpretability and maintaining reasonable predictive accuracy. However, the uneven compliance across borrower segments highlights the need for a more adaptive approach—one that can tailor model behavior to different income groups. The next modeling strategy addresses this need by introducing a segment-aware objective function that allows for differentiated treatment of borrowers based on their risk profiles and regulatory sensitivity.

### Segment-Aware Huber with Threshold Penalty

Building on the limitations identified in previous models, this strategy introduces a segment-aware extension to improve compliance consistency across income groups. Although both the post-hoc quantile calibration and the initial pre-hoc Huber + threshold objective successfully reduced overall threshold violations, they failed to achieve uniform performance. The low-income segment, in particular, continued to exceed the 10% regulatory threshold—highlighting the need for a more targeted approach.

The segment-aware objective addresses this by varying the model’s behavior across income groups. Specifically, it introduces group-specific values for the Huber transition point and the penalty weight , allowing the loss function to adjust both error sensitivity and regulatory conservatism based on the borrower’s segment.

The overall loss is composed of two terms. The first is the **Huber loss**, which penalizes prediction errors while being robust to outliers. In the segment-aware version, the transition point  varies by segment:

The second component is the **regulatory penalty term**, which activates only when a prediction exceeds the allowable threshold:

Here, reflects how strictly the model should penalize regulatory violations within each group. A higher ​ leads to more conservative predictions.

The complete segment-aware loss function combines both components:

This formulation allows the model to internalize both predictive accuracy and compliance constraints in a segment-sensitive manner. By adapting its behavior to the risk characteristics of each income group, the model aims to deliver more balanced and equitable performance across the population.

The segment-specific parameters were selected through experimentation on the training set to balance accuracy and compliance across groups:

* Low-income segment (≤ 1500 GEL):
  + ,
  + Enforces tight error tolerance and strong penalization for overestimation, reflecting high regulatory sensitivity.
* Mid-income segment (1500–2500 GEL)
  + ,
  + Applies moderate tolerance and penalty.
* High-income segment (> 2500 GEL)
  + ,
  + Allows greater tolerance and minimal penalty, given the lower supervisory risk in this group.

Final evaluation was performed on a separate, unseen test set to ensure unbiased performance assessment.

**Results**

The segment-aware Huber + threshold model delivered strong results across all key evaluation metrics. On the test set, it achieved an RMSE of 1227.07, an R² of 0.587, and a threshold exceedance rate of 7.60%, the lowest of all models tested. These results indicate that the model effectively balances prediction accuracy with regulatory compliance.

A notable strength of this model is its ability to maintain stable and fair performance across different income segments. As shown in Figure 7 (bottom left), the exceedance rate in the low-income group was reduced to just above 10%, while mid- and high-income groups remained within acceptable ranges. This balance reflects the model’s use of segment-specific penalties and error tolerances, allowing it to adjust its behavior based on the regulatory sensitivity of each group. Instead of applying the same rules to all borrowers, the model responds to differences in risk and income structure—leading to more consistent outcomes across the population.

A close-up of a graph

AI-generated content may be incorrect.

Figure 7: Analysis of the segment-aware pre-hoc custom objective in comparison with baseline and post-hoc models.

In terms of interpretability, the model remains transparent and reliable. Figure 8 shows the top 10 most important features used by the segment-aware model. Variables such as Min\_transfer\_In, Inc\_6M, and Inc\_in rank highest, confirming that the model bases its predictions on stable and economically meaningful inputs. The use of familiar financial indicators suggests that the addition of segment-specific penalties did not distort the model’s internal logic, preserving its clarity and trustworthiness.

A graph with green bars

AI-generated content may be incorrect.

Figure 8: Top 10 feature importance rankings for the segment-aware pre-hoc model using the Huber + threshold loss.

### Comparison of Pre-Hoc Objective Functions

While both pre-hoc strategies aim to reduce regulatory threshold violations, their outcomes reveal important differences. As shown in **Figure 9 (bottom right)**, the **radar chart summarizes key evaluation metrics on a unified scale where larger values indicate better performance** (for metrics where lower values are preferred an inverse transformation was applied to ensure consistency). The **segment-aware model outperforms the uniform version across all performance and compliance metrics, except the threshold constraint violations the middle- and high-income groups**, where the uniform model shows slightly lower violation rates.

However, this does not imply inferior performance. On the contrary, as shown **in Figure 9 (bottom left)**, it reflects a **more balanced and fair distribution of compliance**. The segment-aware model maintains threshold exceedance rates consistently near the 10% target across all income levels, whereas the uniform model overcorrects in higher segments while failing in the low-income group. This makes the segment-aware approach more effective in practice, as it promotes fairness and aligns better with regulatory expectations.

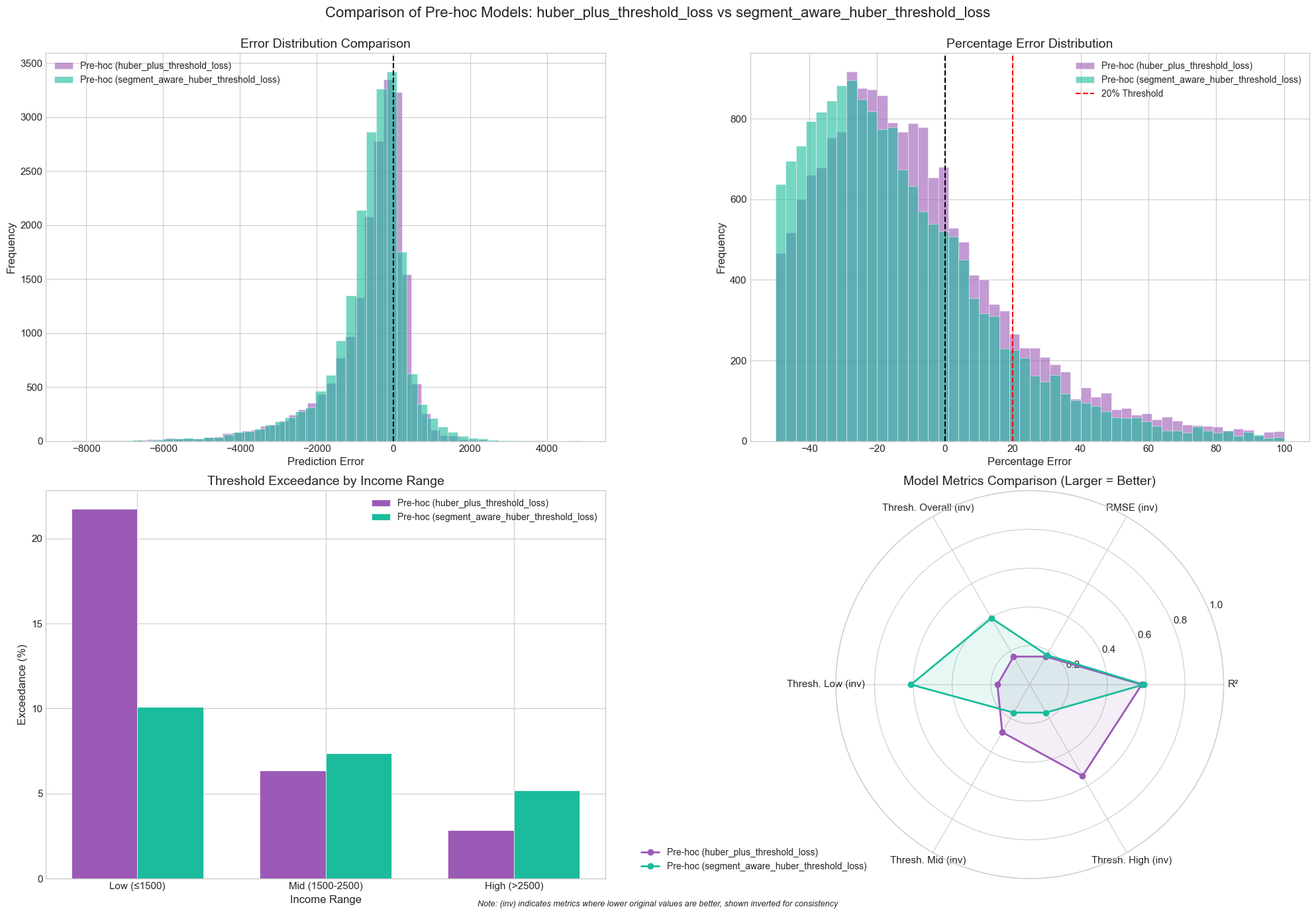


Figure 9: Comparative analysis of the two pre-hoc objectives: huber\_plus\_threshold\_loss vs. segment\_aware\_huber\_threshold\_loss.

Overall, the segment-aware Huber + threshold objective emerges as the more effective and balanced approach. It delivers solid predictive accuracy, regulatory compliance, and equitable treatment across segments—making it the most robust and policy-aligned strategy evaluated in this study.

# Limitations and Directions for Future Work

This thesis has focused on exploring and comparing various modeling approaches for income estimation under regulatory constraints, with the primary goal of evaluating how different objective functions perform in balancing accuracy and supervisory compliance. While the study offers several insights into model behavior, it is important to clarify its scope and acknowledge key limitations.

First, the analysis was conducted entirely on a synthetic dataset. Although synthetic data was generated using CTGAN method, it still may not fully replicate the complexity, correlation structures, or behavior of real-world financial data. Some records contain implausible or internally inconsistent values, which could affect both model learning and evaluation. As a result, the empirical findings in this thesis should be interpreted as illustrative rather than directly generalizable to operational settings. Validating the proposed strategies on real borrower data is crucial to ensure their applicability beyond experimental settings.

Second, the models developed in this thesis were constructed for research purposes and not intended for immediate deployment in production environments. As such, several considerations that are critical for real-world systems — such as stress-testing under adverse scenarios, temporal validation across multiple economic cycles, and performance monitoring over time — were intentionally excluded. These aspects are vital in applied settings to ensure that model outputs remain reliable, fair, and compliant across varying conditions. However, given the scope of this thesis, the priority was placed on methodological experimentation and comparative analysis, rather than full-scale operational robustness.

**Third**, the thesis implicitly assumes that commercial banks aim to optimize model performance — e.g., through improved accuracy. In practice, however, a bank’s primary objective is profitability, and model outputs are only one part of broader portfolio-level decision-making. A more comprehensive evaluation would require simulating lending decisions based on each model’s predictions and analyzing the resulting portfolios in terms of size, composition, credit risk, and financial performance. In the absence of real-world portfolio and loan performance data, this thesis uses traditional performance metrics as a proxy for institutional objectives. While this simplified setup is reasonable for methodological comparison, it limits the ability to draw conclusions about real-world profitability and underscores an important area for future research.

Despite these limitations, the thesis provides a strong foundation for advancing regulatory-sensitive income modeling, particularly in the context of Georgia’s evolving financial landscape. Given the inefficiencies and exclusionary nature of traditional verification methods, the shift toward data-driven models presents a more scalable and inclusive solution. As access to higher-quality, real-world borrower data improves and as institutional objectives become more clearly defined, the strategies developed in this research can support more robust and operationally relevant modeling. The thesis shows that by integrating supervisory priorities into model design — particularly through population-aware penalty structures — it is possible to address both accuracy and compliance concerns in a structured way. Taken together, these contributions lay the groundwork for continued exploration at the intersection of statistical modeling, credit risk assessment, and regulatory compliance in Georgia’s financial sector.

# Bibliography

Kibekbaev, A., & Duman, E. (2016). Benchmarking Regression Algorithms for Income Prediction Modeling. *Information Systems,, 61*, 40-52.

Lazar, A. (2004). Income prediction via support vector machine. *Proceedings of the 2004 International Conference on Machine Learning and Applications* (გვ. 143–149). Louisville, KY, USA: IEEE.

Farrell, D., & Greig, F. (2018). *Estimating family income from administrative banking data.* JPMorgan Chase Institute.

Suarez, G., Raful, J., Luque, M. A., Valencia, C. F., & Correa-Bahnsen, A. (2021). *Enhancing user's income estimation with super-app alternative data.* arXiv: https://arxiv.org/abs/2104.05831

Matkowski, M. (2021). *Prediction of Individual Level Income: A Machine Learning Approach.* Bryant Digital Repository: https://digitalcommons.bryant.edu/honors\_economics/39/

Kokhreidze, G. (2025). *Banking Sector Clients Income Estimation Models.* Tbilisi: TSU.

National Bank of Georgia. (2020). *Regulation about Data-driven Statistical, AI, and Machine Learning Model Risk Management.* https://nbg.gov.ge/fm/ზედამხედველობა/ფინტექ\_დეპარტამენტი/მოდელ\_რისკი/mrm-151-04eng.pdf

SARGAN, Ç. (2021). *Income Estimation Model for Individual Customers.* https://openaccess.bilgi.edu.tr/server/api/core/bitstreams/975d49c3-6786-4021-97d6-84fec9844827/content

# Appendix A: Computational Source – Jupyter Notebook

This thesis is supported by a Jupyter notebook titled notebook.ipynb, which contains the full implementation of the analytical workflow. It includes all preprocessing procedures, data exploration, model development, calibration techniques, and evaluation outputs referenced throughout the thesis.

The notebook is organized to mirror the thesis structure, covering:

* Setup and Data Loading
* Exploratory Data Analysis
  + Missing Value Analysis
  + Statistical Summary
* Data Visualization and Analysis
  + Target Distribution
  + Univariate and Multivariate Analysis
* Data Preparation
  + Loading, Preprocessing Strategy, and Implementation
* Modeling
  + Baseline XGBoost Model
  + Post-hoc Calibrated XGBoost Model
    - Quantile Regression Calibration
  + Pre-hoc Adjusted Models
    - Huber Loss with Threshold Penalty
    - Segment-Aware Huber with Threshold Penalty
  + Model Comparison and Evaluation

All figures and numerical findings in this thesis are derived directly from this source. The notebook is available [here](https://github.com/Sandro-Gogaladze/BA_Thesis.git) for verification and reproducibility.